An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Make a Big Difference?

Ryan Giordano (rgiordan@mit.edu)[1]
January 2022

[1]With coauthors Rachael Meager (LSE) and Tamara Broderick (MIT)

An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Make a Big Difference?

The conclusions of one's statistical analysis may depend on only a **small fraction of the data**, even for **highly significant results in correctly specified models**.

Ryan Giordano (rgiordan@mit.edu)[1]
January 2022

[1]With coauthors Rachael Meager (LSE) and Tamara Broderick (MIT)

An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Make a Big Difference?

The conclusions of one's statistical analysis may depend on only a **small fraction of the data**, even for **highly significant results in correctly specified models**.

We provide a **generally applicable tool** to detect such sensitivity. Our methods are **efficiently and automatically computable**, and come with **finite-sample accuracy guarantees** and **clear intuition**.

Ryan Giordano (rgiordan@mit.edu)[1]
January 2022

[1]With coauthors Rachael Meager (LSE) and Tamara Broderick (MIT)

Dropping data: Mexico Microcredit

Example: Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Beta (SE)
Original result	-4.55 (5.88)

The original conclusion: No evidence that microcredit is effective...

$\Rightarrow$ Standard errors can be inadequate summaries of data sensitivity!

Cannot find influential subsets by brute force! $\binom{16{,}560}{15} \approx 1.5 \cdot 10^{51}$

We provide a fast, automatic tool to approximately identify the most influential set of points.

Dropping data: Mexico Microcredit

Example: Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)

The original conclusion: No evidence that microcredit is effective...

$\Rightarrow$ Standard errors can be inadequate summaries of data sensitivity!

Cannot find influential subsets by brute force! $\binom{16{,}560}{15} \approx 1.5 \cdot 10^{51}$

We provide a fast, automatic tool to approximately identify the most influential set of points.

Dropping data: Mexico Microcredit

Example: Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
"Significant" negative change	14	-10.96 (5.57)

The original conclusion: No evidence that microcredit is effective...

$\Rightarrow$ Standard errors can be inadequate summaries of data sensitivity!

Cannot find influential subsets by brute force! $\binom{16{,}560}{15} \approx 1.5 \cdot 10^{51}$

We provide a fast, automatic tool to approximately identify the most influential set of points.

Dropping data: Mexico Microcredit

Example: Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
"Significant" negative change	14	-10.96 (5.57)
"Significant" positive change	15	7.03 (2.55)

The original conclusion: No evidence that microcredit is effective...

$\Rightarrow$ Standard errors can be inadequate summaries of data sensitivity!

Cannot find influential subsets by brute force! $\binom{16{,}560}{15} \approx 1.5 \cdot 10^{51}$

We provide a fast, automatic tool to approximately identify the most influential set of points.

Dropping data: Mexico Microcredit

Example: Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
"Significant" negative change	14	-10.96 (5.57)
"Significant" positive change	15	7.03 (2.55)

The original conclusion: No evidence that microcredit is effective...
...can be reversed by dropping less than 0.1% of the data.

$\Rightarrow$ Standard errors can be inadequate summaries of data sensitivity!

Cannot find influential subsets by brute force! $\binom{16,560}{15} \approx 1.5 \cdot 10^{51}$

We provide a fast, automatic tool to approximately identify the most influential set of points.

Outline

- Why and when might you care about sensitivity to data dropping?
- How does our approximation work, and when is it accurate?

  (A formalization of the problem and the class of estimators we study.)
- Examine real-life examples of analyses: some sensitive, some not.

  (The results may defy your intuition.)
- What kinds of analyses are sensitive to data dropping?

  (Including comparison to standard errors and gross-error robustness.)

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Not always! But sometimes, surely yes.

Thinking without random noise can be helpful.

Suppose you have a farm, and want to know whether your average yield is greater than 170 bushels per acre. At harvest, you measure 200 bushels per acre.

- Scenario one: If your yield is greater than 170 bushels per acre, you make a profit.
  - Don't care about sensitivity to small subsets
- Scenario two: You want to recommend your farming methods to a friend across the valley.
  - Might care about sensitivity to small subsets

For example, often in economics:

- Small fractions of data are missing not-at-random,
- Policy population is different from analyzed population,
- We report a convenient summary (e.g. mean) of a complex effect,
- Models are stylized proxies of reality.

Formalizing the question.

Ordinary least squares

A data point d_n has regressors x_n and response y_n: $d_n = (x_n, y_n)$.

The estimator $\hat{\theta} \in \mathbb{R}^P$ satisfies:

$$\hat{\theta} := \arg\min_{\theta} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \theta^T x_n\right)^2$$

$$\Leftrightarrow \sum_{n=1}^{N} \left(y_n - \hat{\theta}^T x_n\right) x_n = 0.$$

Make a qualitative decision using:

- A particular component: θ_k
- The end of a confidence interval: $\theta_k + \frac{1.96}{\sqrt{N}} \hat{\sigma}(\hat{\theta})$

Z-estimators

We observe N data points $d_1, \dots, d_N$ (in any domain).

The estimator $\hat{\theta} \in \mathbb{R}^P$ satisfies:

$$\sum_{n=1}^{N} G(\hat{\theta}, d_n) = 0_P.$$

$G(\cdot, d_n)$ is "nice," $\mathbb{R}^P$-valued. E.g. OLS, MLE, VB, IV &c.

Make a qualitative decision using $\phi(\hat{\theta})$ for a smooth, real-valued ϕ.

(WLOG try to increase $\phi(\hat{\theta})$.)

Question: Can we make a big change in $\phi(\hat{\theta})$ by dropping $\lfloor \alpha N \rfloor$ datapoints, for some small proportion α?

Which estimators do we study?

Question: Can we make a big change in $\phi(\hat{\theta})$ by dropping $\lfloor \alpha N \rfloor$ datapoints, for some small proportion α? **Two big problems:**

- There are $\binom{N}{\lfloor \alpha N \rfloor}$ sets to check. (Huge even for $\alpha \ll 1$.)
- Evaluating $\hat{\theta}$ re-solving the estimating equation.
  - E.g., re-computing the OLS estimator.
  - Other examples are even harder (VB, machine learning)

Idea: Smoothly approximate the effect of leaving out points.

We have N data points $d_1, \ldots, d_N$, a quantity of interest $\phi(\cdot)$, and

$$\sum_{n=1}^{N} G(\hat{\theta}, d_n) = 0_P \qquad .$$

Which estimators do we study?

Question: Can we make a big change in $\phi(\hat{\theta})$ by dropping $\lfloor \alpha N \rfloor$ datapoints, for some small proportion α? **Two big problems:**

- There are $\binom{N}{\lfloor \alpha N \rfloor}$ sets to check. (Huge even for $\alpha \ll 1$.)
- Evaluating $\hat{\theta}$ re-solving the estimating equation.
  - E.g., re-computing the OLS estimator.
  - Other examples are even harder (VB, machine learning)

Idea: Smoothly approximate the effect of leaving out points.

We have N data points $d_1, \ldots, d_N$, a quantity of interest $\phi(\cdot)$, and

$$\sum_{n=1}^{N} \vec{w}_n G(\hat{\theta}(\vec{w}), d_n) = 0_P \text{ for a weight vector } \vec{w} \in \mathbb{R}^N.$$

Which estimators do we study?

Question: Can we make a big change in $\phi(\hat{\theta})$ by dropping $\lfloor \alpha N \rfloor$ datapoints, for some small proportion α? **Two big problems:**

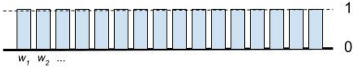
- There are $\binom{N}{\lfloor \alpha N \rfloor}$ sets to check. (Huge even for $\alpha \ll 1$.)
- Evaluating $\hat{\theta}$ re-solving the estimating equation.
    - E.g., re-computing the OLS estimator.
    - Other examples are even harder (VB, machine learning)

Idea: Smoothly approximate the effect of leaving out points.

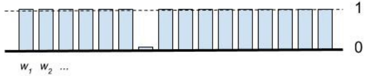
We have N data points $d_1, \ldots, d_N$, a quantity of interest $\phi(\cdot)$, and

$$\sum_{n=1}^{N} \vec{w}_n G(\hat{\theta}(\vec{w}), d_n) = 0_P \text{ for a weight vector } \vec{w} \in \mathbb{R}^N.$$

Original weights: $\vec{1} = (1, \ldots, 1)$



Leave points out by setting their elements of $\vec{w}$ to zero.

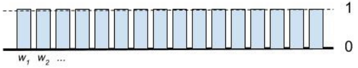


The map $\vec{w} \mapsto \phi(\hat{\theta}(\vec{w}))$ is well-defined even for continuous weights.

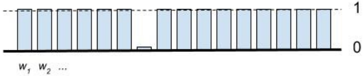
Which estimators do we study?

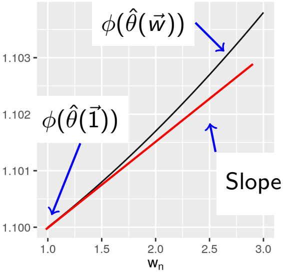
$$\sum_{n=1}^{N} \vec{w}_n G(\hat{\theta}(\vec{w}), d_n) = 0_P \text{ for a weight vector } \vec{w} \in \mathbb{R}^N.$$

Original weights: $\vec{1} = (1, \ldots, 1)$



Leave points out by setting their elements of $\vec{w}$ to zero.





Slope $= \left.\frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n}\right|_{\vec{1}} =: \psi_n$

The slopes ψ_n are the **empirical influence function** [Hampel, 1986]. We call them "influence scores."

We can use ψ_n to form a Taylor series approximation:

$$\phi(\hat{\theta}(\vec{w})) \approx \color{red}{\phi^{\mathrm{lin}}(\vec{w})} := \phi(\hat{\theta}(\vec{1})) + \sum_{n=1}^{N} \psi_n(\vec{w}_n - 1)$$

Taylor series approximation.

Problem: How much can you change $\phi(\hat{\theta}(\vec{w}))$ dropping $\lfloor \alpha N \rfloor$ points?
Combinatorially hard by brute force!

Approximate Problem: How much can you change $\phi^{\mathrm{lin}}(\hat{\theta}(\vec{w}))$ dropping $\lfloor \alpha N \rfloor$ points? **Easy!**

$$\phi^{\mathrm{lin}}(\vec{w}) := \phi(\hat{\theta}(\vec{1})) + \sum_{n=1}^{N} \psi_n (\vec{w}_n - 1)$$

Dropped points have $\vec{w}_n - 1 = -1$. Kept points have $\vec{w}_n - 1 = 0$
$\Rightarrow$ The most influential points for $\phi^{\mathrm{lin}}(\vec{w})$ have the most negative ψ_n.

Procedure:

1. Compute your original estimator $\hat{\theta}(\vec{1})$.
2. Compute and sort the influence scores $\psi_{(1)}, \ldots, \psi_{(N)}$.
3. Worry if $-\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)}$ is large enough to change your conclusions.

How to compute the influence scores?

How can we compute the influence scores $\psi_n = \left. \frac{\partial \phi(\hat{\theta}(\vec{w}))}{\partial \vec{w}_n} \right|_{\vec{1}}$?

By the **chain rule**, $\psi_n = \left. \frac{\partial \phi(\theta)}{\partial \theta} \right|_{\hat{\theta}(\vec{1})} \left. \frac{\partial \hat{\theta}(\vec{w})}{\partial \vec{w}_n} \right|_{\vec{1}}$.

Recall that $\sum_{n=1}^{N} \vec{w}_n G(\hat{\theta}(\vec{w}), d_n) = 0_P$ for all $\vec{w}$ near $\vec{1}$.

$\Rightarrow$ By the **implicit function theorem**, we can write $\left. \frac{\hat{\theta}(\vec{w})}{\partial \vec{w}_n} \right|_{\vec{1}}$ as a linear system involving $G(\cdot, \cdot)$ and its derivatives.

$\Rightarrow$ The ψ_n are automatically computable from $\hat{\theta}(\vec{1})$ and software implementations of $G(\cdot, \cdot)$ and $\phi(\cdot)$ using **automatic differentiation**.

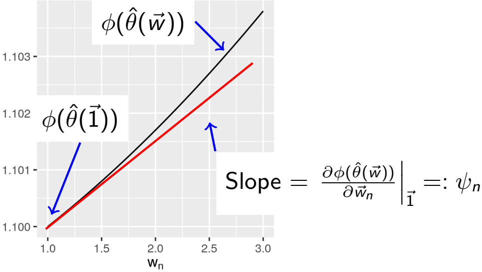
```
import jax
import jax.numpy as np
def phi(theta):
    ... computations using np and theta ...
    return value

# Exact gradient of phi (1st term in the chain rule):
jax.grad(phi)(theta_opt)
```

See `rgiordan/vittles` and `rgiordan/zaminfluence` on `github`.

How accurate is the approximation?

By conrolling the curvature, we can control the error in the linear approximation.



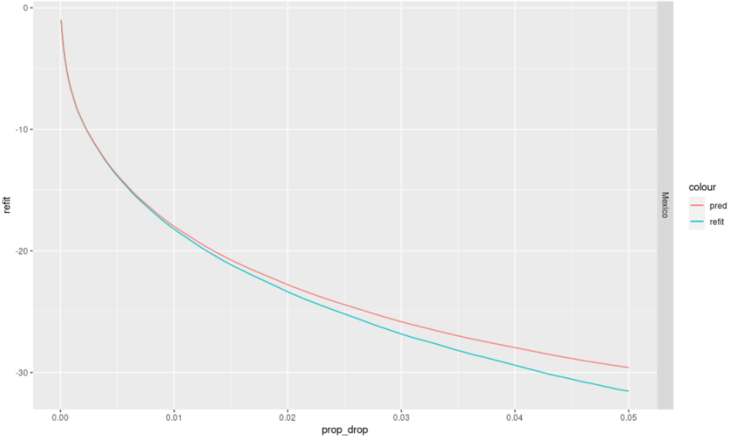
We provide **finite-sample theory** showing that

$$\left|\phi(\hat{\theta}(\vec{w})) - \phi^{\mathrm{lin}}(\vec{w})\right| = O\left(\left\|\frac{1}{N}(\vec{w} - \vec{1})\right\|_2^2\right) = O(\alpha) \text{ as } \alpha \to 0.$$

You don't need to rely on the theory!

Our method returns the set of points to drop. Re-running once without those points provides an **exact lower bound** on the true worst-case sensitivity.

Mexico example:
See `microcredit_profit_sandbox.R`.



Selected experimental results.

Study case	Original estimate (SE)	Target change	Refit estimate	Observations dropped
Mexico	-4.549 (5.879)	Sign change	**0.398 (3.194)**	$1 = 0.01\%$
		Significance change	-**10.962 (5.565)***	$14 = 0.08\%$
		Significant sign change	**7.030 (2.549)***	$15 = 0.09\%$

Table: Microcredit Mexico results [Angelucci et al., 2015].

Selected experimental results.

Study case	Original estimate (SE)	Target change	Refit estimate	Observations dropped
Mexico	-4.549 (5.879)	Sign change	**0.398 (3.194)**	1 = 0.01%
		Significance change	-**10.962 (5.565)***	14 = 0.08%
		Significant sign change	**7.030 (2.549)***	15 = 0.09%

Table: Microcredit Mexico results [Angelucci et al., 2015].

Study case	Original estimate (SE)	Target change	Refit estimate	Observations dropped
Poor, period 10	33.861 (4.468)*	Sign change	-**2.559 (3.541)**	697 = 6.63%
		Significance change	**4.806 (3.684)**	435 = 4.14%
		Significant sign change	-**9.416 (3.296)***	986 = 9.37%

Table: Cash transfers results. [Angelucci and De Giorgi, 2009]

Selected experimental results.

Study case	Original estimate (SE)	Target change	Refit estimate	Observations dropped
Mexico	-4.549 (5.879)	Sign change	**0.398 (3.194)**	1 = 0.01%
		Significance change	-**10.962 (5.565)***	14 = 0.08%
		Significant sign change	**7.030 (2.549)***	15 = 0.09%

Table: Microcredit Mexico results [Angelucci et al., 2015].

Study case	Original estimate (SE)	Target change	Refit estimate	Observations dropped
Poor, period 10	33.861 (4.468)*	Sign change	-**2.559 (3.541)**	697 = 6.63%
		Significance change	**4.806 (3.684)**	435 = 4.14%
		Significant sign change	-**9.416 (3.296)***	986 = 9.37%

Table: Cash transfers results. [Angelucci and De Giorgi, 2009]

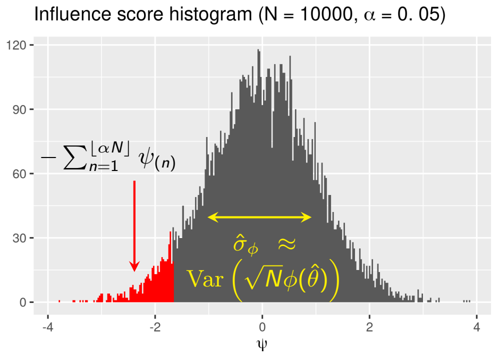
Study case	Original estimate (SE)	Target change	Refit estimate	Observations dropped
Health notpoor 12m	0.029 (0.005)*	Sign change	-**0.001 (0.005)**	156 = 0.67%
		Significance change	**0.008 (0.005)**	101 = 0.43%
		Significant sign change	-**0.009 (0.004)***	224 = 0.96%

Table: Medicaid profit results [Finkelstein et al., 2012]

What makes an estimator non-robust? A tail sum.

We show that $\phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = -\sum_{n=1}^{\lfloor \alpha N \rfloor} \psi_{(n)} =: \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha$ where

- The "noise" $\hat{\sigma}_\phi^2 \to \mathrm{Var}(\sqrt{N}\phi)$
  - $\hat{\sigma}_\phi^2 =$ is the robust "sandwich" variance estimator [Hampel, 1986]
- The "shape" $\hat{\mathscr{T}}_\alpha \le \sqrt{\alpha(1-\alpha)}$ determined by ψ_n distribution
  - $\hat{\mathscr{T}}_\alpha$ converges to a nonzero constant



Example.

Report non-robustness if:

$$\phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \geq \Delta \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \leq \hat{\mathscr{T}}_\alpha.$$

The **signal to noise ratio** $\frac{\Delta}{\hat{\sigma}_\phi}$ determines sensitivity to data dropping.

Study case	Original estimate (SE)	Target change	Refit estimate	Observations dropped
Health notpoor 12m	0.029 (0.005)*	Sign change	**-0.001 (0.005)**	156 = 0.67%
		Significance change	**0.008 (0.005)**	101 = 0.43%
		Significant sign change	**-0.009 (0.004)***	224 = 0.96%

Table: Medicaid profit results [Finkelstein et al., 2012]

Let's analyze with $\alpha = 0.01 = 1\%$.

$$\begin{aligned}
\phi(\hat{\theta}) &= -0.029 && \text{(Increase QOI by defn)} & \Delta &= 0.029 \\
\hat{\sigma}_\phi &= 0.766 && \text{(Noise)} & \frac{1}{\sqrt{N}}\hat{\sigma}_\phi &= 0.005 && \text{(SE)} \\
\hat{\mathscr{T}}_\alpha &= 0.046 && \text{(Shape)} & \frac{1.96}{\sqrt{N}} &= 0.0128 && \to 0 \text{ as } N \to \infty \\
\hat{\mathscr{T}}_\alpha \hat{\sigma}_\phi &= 0.035 && \text{(Data dropping sensitivity)} & \frac{1.96}{\sqrt{N}}\hat{\sigma}_\phi &= 0.010 && \text{(SE sensitivity)}
\end{aligned}$$

The noise is much larger than the signal $\Rightarrow$ Sensitive to data dropping.

Corollaries.

Report non-robustness if:

$$\phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \ge \Delta \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \le \hat{\mathscr{T}}_\alpha.$$

The **signal to noise ratio** $\frac{\Delta}{\hat{\sigma}_\phi}$ determines sensitivity to data dropping.

Corollaries.

Report non-robustness if:

$$\phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \geq \Delta \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \leq \hat{\mathscr{T}}_\alpha.$$

The **signal to noise ratio** $\frac{\Delta}{\hat{\sigma}_\phi}$ determines sensitivity to data dropping.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.
Standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \leq \frac{1.96}{\sqrt{N}} \neq \hat{\mathscr{T}}_\alpha$.

Corollaries.

Report non-robustness if:

$$\phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \ge \Delta \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \le \hat{\mathscr{T}}_\alpha.$$

The **signal to noise ratio** $\frac{\Delta}{\hat{\sigma}_\phi}$ determines sensitivity to data dropping.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.
Standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \le \frac{1.96}{\sqrt{N}} \ne \hat{\mathscr{T}}_\alpha$.

Corollary: Statistical insignificance is asymptotically non-robust.
$\frac{1.96\hat{\sigma}_\phi}{\sqrt{N}} \to 0 \le \hat{\mathscr{T}}_\alpha$.

Corollaries.

Report non-robustness if:

$$\phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \geq \Delta \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \leq \hat{\mathscr{T}}_\alpha.$$

The **signal to noise ratio** $\frac{\Delta}{\hat{\sigma}_\phi}$ determines sensitivity to data dropping.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.
Standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \leq \frac{1.96}{\sqrt{N}} \neq \hat{\mathscr{T}}_\alpha$.

Corollary: Statistical insignificance is asymptotically non-robust.
$\frac{1.96\hat{\sigma}_\phi}{\sqrt{N}} \to 0 \leq \hat{\mathscr{T}}_\alpha$.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.
Both $\hat{\mathscr{T}}_\alpha$ and $\hat{\sigma}_\phi$ typically converge to nonzero constants.

Corollaries.

Report non-robustness if:

$$\phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \ge \Delta \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \le \hat{\mathscr{T}}_\alpha.$$

The **signal to noise ratio** $\frac{\Delta}{\hat{\sigma}_\phi}$ determines sensitivity to data dropping.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.
Standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \le \frac{1.96}{\sqrt{N}} \ne \hat{\mathscr{T}}_\alpha$.

Corollary: Statistical insignificance is asymptotically non-robust.
$\frac{1.96\hat{\sigma}_\phi}{\sqrt{N}} \to 0 \le \hat{\mathscr{T}}_\alpha$.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.
Both $\hat{\mathscr{T}}_\alpha$ and $\hat{\sigma}_\phi$ typically converge to nonzero constants.

Corollary: Non-robustness possible even with correct specification.

Corollaries.

Report non-robustness if:

$$\phi^{\mathrm{lin}}(\vec{w}^*) - \phi(\hat{\theta}) = \hat{\sigma}_\phi \hat{\mathscr{T}}_\alpha \ge \Delta \qquad \Leftrightarrow \qquad \frac{\Delta}{\hat{\sigma}_\phi} \le \hat{\mathscr{T}}_\alpha.$$

The **signal to noise ratio** $\frac{\Delta}{\hat{\sigma}_\phi}$ determines sensitivity to data dropping.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out is different from standard errors.
Standard errors reject when $\frac{\Delta}{\hat{\sigma}_\phi} \le \frac{1.96}{\sqrt{N}} \ne \hat{\mathscr{T}}_\alpha$.

Corollary: Statistical insignificance is asymptotically non-robust.
$\frac{1.96\hat{\sigma}_\phi}{\sqrt{N}} \to 0 \le \hat{\mathscr{T}}_\alpha$.

Corollary: Leave-$\lfloor \alpha N \rfloor$-out robustness does not vanish as $N \to \infty$.
Both $\hat{\mathscr{T}}_\alpha$ and $\hat{\sigma}_\phi$ typically converge to nonzero constants.

Corollary: Non-robustness possible even with correct specification.

Corollary: To robustify, reduce the noise or increase the signal.

Other forms of robustness

We proceeded as follows:

1. Took presence of datapoints as a model input,
2. Formed an automatically-computable differential approximation,
3. Provided theory by analyzing higher-order derivatives,
4. Studied its effectiveness in problems with open-access data.

Presence of datapoints is only one model input of many!

- Prior sensitivity in Bayesian nonparametrics [Giordano et al., 2021]
- Model sensitivity of MCMC output [Gustafson, 2000, Giordano et al., 2018]
- Cross-validation [Giordano et al., 2019, Wilson et al., 2020]
- Cross-validation [Giordano et al., 2019, Wilson et al., 2020]
- Frequentist variances of MCMC posteriors (in progress)

Conclusion

- You may be concerned if you could reverse your conclusion by removing a small proportion of your data.

Conclusion

- You may be concerned if you could reverse your conclusion by removing a small proportion of your data.
- We can quickly and automatically find an approximate influential set which is accurate for small sets.

Conclusion

- You may be concerned if you could reverse your conclusion by removing a small proportion of your data.
- We can quickly and automatically find an approximate influential set which is accurate for small sets.
- Robustness to removing small sets is principally determined by the signal to noise ratio.

Conclusion

- You may be concerned if you could reverse your conclusion by removing a small proportion of your data.
- We can quickly and automatically find an approximate influential set which is accurate for small sets.
- Robustness to removing small sets is principally determined by the signal to noise ratio.
- In the present work, we studied data dropping. But we provide a framework for studying many other robustness questions, both to data and model perturbations.

Links and references

Tamara Broderick, Ryan Giordano, Rachael Meager (alphabetical authors)
“An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?”
https://arxiv.org/abs/2011.14999

Open-source software:

R package `zaminfluence`	`https://github.com/rgiordan/zaminfluence`
Python package `vittles`	`https://github.com/rgiordan/vittles`

Some related content can be found on my blog:
`https://rgiordan.github.io/`

References

M. Angelucci and G. De Giorgi. Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *American Economic Review*, 99(1):486–508, 2009.

M. Angelucci, D. Karlan, and J. Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82, 2015.

A. Baydin, B. Pearlmutter, A. Radul, and J. Siskind. Automatic differentiation in machine learning: A survey. *The Journal of Machine Learning Research*, 18(1):5595–5637, 2017.

A. Finkelstein, S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. Newhouse, H. Allen, K. Baicker, and Oregon Health Study Group. The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106, 2012.

R. Giordano, T. Broderick, and M. I. Jordan. Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029, 2018.

R. Giordano, W. Stephenson, R. Liu, M. I. Jordan, and T. Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR, 2019.

R. Giordano, R. Liu, M. I. Jordan, and T. Broderick. Evaluating sensitivity to the stick-breaking prior in Bayesian nonparametrics. 2021.

P. Gustafson. Local robustness in Bayesian analysis. In *Robust Bayesian Analysis*, pages 71–88. Springer, 2000.

F. Hampel. *Robust statistics: The approach based on influence functions*, volume 196. Wiley-Interscience, 1986.

A. Wilson, M. Kasy, and L. Mackey. Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4530–4540. PMLR, 2020.